RAVI KIRAN VADLAMANI

412 251 3366 | kiranvadlamani94@gmail.com | www.linkedin.com/in/ravi-kiran-v

SUMMARY

Software Engineer with 4+ years building large-scale distributed systems. Currently SDE II at AWS shipping high availability, performance, and correctness improvements to a Tier-1 real-time service. Prior experience in production ML serving, data pipelines, and high-throughput backends.

SKILLS

Programming Languages: Java,, Javascript, Python, C/C++
Frameworks: Spring (Core, Boot, Security, Data), Hibernate, JPA, Maven, Gradle, FastAPI, Flask, Django, React, Angular, Redux, Node.js, Express, HTML, CSS, SASS, Kafka, Git, Agile/Scrum, CI/CD, Unit Testing (Junit, Jest, Mocha), GraphQL, TensorRT-LM, Triton Inference Server
Cloud & Devops: AWS (DynamoDB, CloudWatch, Lambda, CDK), Docker, Kubernetes
Systems: Distributed consensus, leader election, DynamoDB, sharding, caching, replication, event-driven architectures, load testing, observability

INDUSTRIAL EXPERIENCE

Amazon Web Services (AWS), Bellevue, WA ***Mar 2025 - Present***
Software Development Engineer II
Tech: Java 17, Spring, DynamoDB, AWS CDK, CloudWatch, SNS/SQS
● Improved P99 latency of a real-time routing hot path by ~70% through a parallelized execution redesign for the largest-scale customer tenants; validated via load testing that exceeded peak production traffic.
● Delivered an upstream data-shaping optimization that reduced search-index query time by ~83% with no correctness regressions.
● Authored high- and low-level designs for a distributed-locking redesign that activates all hosts in a cluster (vs. a single leader), reducing single-host-failure blast radius by ~75% and improving resource utilization.
● Built end-to-end canary infrastructure (CDK + integration test framework) with automated deployment gating — regressions in a composite set of alarms block pipeline promotion.
● Built internal GenAI tooling to automate operational review workflows and on-call gameday procedures; recognized with a team-level AI innovation award.
● Active code reviewer — caught runtime null-pointer bugs, architectural coupling issues, and infrastructure-as-code risks in peer changes pre-merge.

Apexon, Santa Clara, CA ***Feb 2023 - Mar 2025***
Software Engineer II
Languages & Technologies: Java(Spring), Python(Flask), Snowflake, openshift, control-m
● Optimized a Data Quality profiler via Python multithreading and SQL query tuning, reducing runtime from 4 hours to ~30 minutes (~8x).
● Built Flask APIs to validate confidential documents against a standardized schema; owned inference deployment and operational maintenance for production ML models.
● Built the backend framework for generative AI use cases on dbt, coordinating LLM calls, SQL transformations, and downstream data pipelines with retry and backoff handling.

Software Engineer Intern, Amazon Lab 126 ***Jun 2022 – Aug 2022***
Project: Alexa Models Perceptibility Improvement and Dynamic Test Set Creation from Production Utterances for Model Benchmarking
Languages & Technologies: Python, Java, Numpy, Pytorch, DynamoDB
● Built a Java service (Guice/Spring) exposing APIs to create, update, and list dynamic test sets for model benchmarking; reduced customer-request-to-test-set turnaround from 2 weeks to <2 hours.
● Built a Python orchestration layer that provisions an EMR cluster, runs a Spark job, and tears it down — used as the compute backend for the above APIs.
● Improved perceptibility of Alexa model outputs on low-confidence customer utterances via loss-function changes, collaborating with the science team on evaluation.

Indian Oil Corporation Limited, Paradeep, India **Aug 2016 - Aug 2021**
Assistant Manager, Paradip LPG Terminal
Languages & Technologies: Spring Boot, FastAPI, Node.js, Express.js, Spark, AWS
● Led a 10-person cross-functional team (software, automation, security) operating a large-scale LPG terminal; built a gas-leak prediction and interlock shutdown system and a vision-based personnel-tracking system for hazardous zones.
Senior Engineer/Officer, Paradip LPG Terminal
Languages & Technologies: Spring Boot, NodeJS, Spark, AWS
- Designed & developed a robust and scalable distributed web application encompassing features for users and messages that is capable of handling 100,000 concurrent API requests and 1 million database records.

EDUCATION

Carnegie Mellon University, Pittsburgh, PA **Dec-2022**
Master of Science in Electrical and Computer Engineering (Concentration: AI/ML Systems)
National Institute of Technology (NIT), Tiruchirappalli, India **May-2015**
Bachelor of Technology in Instrumentation and Control Engineering